International Journal of Strategic Information Technology and Applications
Volume 8 • Issue 3 • July-September 2017

Yet Another Multidimensional Model for XML Documents

Maha Azabou, Faculty of Economics and Management, University of Sfax, Sfax, Tunisia
Kais Khrouf, Jouf University, Sakakah, Saudi Arabia
Jamel Feki, FCIT, IS Department, University of Jeddah, Jeddah, Saudi Arabia
Chantal Soulé-Dupuy, University of Toulouse 1 Capitole, Toulouse, France
Nathalie Vallès, University of Toulouse 1 Capitole, Toulouse, France

ABSTRACT

The Diamond model is a multidimensional model dedicated to XML document warehouses. It considers structured and unstructured data simultaneously. Furthermore, it orders the semantics of documents via a specific semantic dimension linked to conventional dimensions, thus breaking the classical orthogonality rule of dimensions. After giving an overview of their three-phase quasi-automatic approach for the generation of the diamond model, the authors focus on the Diamond-Gen software tool that supports the proposed approach. The authors illustrate the Diamond-Gen functionalities and assess it through an experimental study using a set of 1500 XML documents issued from the PubMed collection.

KEYWORDS

Diamond model, XML document warehouse, Semantic dimension

1. INTRODUCTION

To support their data analytical processes, today's organizations deploy data warehouses and client OLAP tools (On-Line Analytical Processing) to access, analyze and visualize integrated and aggregated data. The literature review distinguishes two categories of data: structured, and unstructured. Structured data often organize either as relational databases for OLPT (On-Line Transaction Processing) or as data-centric documents for electronic/Internet uses, whereas unstructured data are presented as text-oriented documents. To manage and analyze these both categories, organizations exploit the powerful features the eXtensible Markup Language (XML) offers. XML documents constitute a core source for some decisional analyses since their contents help decision makers to better manage the evolution of business processes. An XML document is generally compliant to a generic grammar called DTD (Document Type Definition). In order to conduct analyses on XML documents it is necessary to extend the traditional data warehouse model (Golfarelli, Maio, & Rizzi, 1998), initially dedicated to numeric data, with capabilities for handling the content of these documents and, in particular, focusing on the document semantics.

To do so, we have proposed the Diamond multidimensional model (Azabou, Khrouf, Feki, Soulé-Dupuy, & Vallès, 2016) dedicated to the design and OLAP of XML documents. The Diamond model focuses on three aspects: (1) text-data structure, (2) text-data semantic, and (3) flexibility of the analysis in the sense the specification of multidimensional analyses performs without constraining

DOI: 10.4018/IJSITA.2017070105

the decision-maker with predefined subjects of analysis (i.e., facts). Thus, the Diamond model offers OLAP analysis on the semantics contained in XML text-oriented documents and on structural data simultaneously. This semantic OLAPing is becoming possible thanks to two new dimensions introduced in the Diamond model, namely Semantic and Standard dimensions (Azabou, Khrouf, Feki, Soulé-Dupuy, & Vallès, 2014).

Furthermore, we help the designer designing a document warehouse (DocW) schema according to the Diamond model starting from a collection of XML documents and the logical structure (DTD or XSchema). For this purpose, we suggest first, a design approach relying on a set of eleven heuristic rules for determining the components of the model (e.g., dimensions and hierarchies), and secondly, we develop a software prototype called Diamond-Gen that supports the rule-based approach; it enables generating automatically a Diamond DocW schema from a set of XML documents.

This paper is organized as follows. Section 2 presents related works dealing with the multidimensional modeling of documents. In Section 3, we give an overview of our approach for building a document warehouse schema compliant to the Diamond model. Section 4 describes our rules for generating a Diamond multidimensional schema adapted to the OLAP analysis of XML documents. Section 5 shows the functionalities of the Diamond-Gen software prototype. Finally, Section 6 concludes the paper and gives an overview of our current works.

2. RELATED WORK

Multidimensional modeling organizes data into data warehouses (DWs) so that OLAP analysis will be easy for decision-makers, efficient and effective. Existing DW models offer a framework for multidimensional modelling of factual data; nevertheless, these models are not appropriate for unstructured textual data. To alleviate such a problem several studies focus on a main issue "How to design a multidimensional model for documents?" One can group current studies into two categories according to the type of used models. The first category suggests extending the conventional multidimensional models (Golfarelli et al., 1998), (Kimball & Ross, 2003), (Immon, 2005) by integrating a semantic dimension, like: 1) Term hierarchy in Text Cube (Lin, Ding, Han, Zhu, & Zhao, 2008).The main idea is to give the user the possibility to make a semantic navigation in data dimension. To specify the semantic level in the text cube, they proposed a hierarchy where the extracted keywords represent the nodes at the base level, the ancestor nodes at upper levels are more general than children at lower level, and the nodes at top level contain terms of the corpus. The use of textual measures pull-up or push down facilitates the navigation in the hierarchy. 2) A topics hierarchy in the Topic Cube (Zhang, Zhai, & Han, 2009) proposal. This structure allows to a user to drill-down and roll-up along this tree and discover the content of the text documents in order to view the different granularities and levels of topics in the cube. The first level in the tree contains the detail of topics, the second level is more general and last level contains the aggregation of all topics. Based on Topic Cube and information network analysis, the authors of (Yu et al., 2009) are interested in automatically constructing concept hierarchies by information network analysis. Such as, NetClus, dealing with multi-typed information network is used for integrated clustering, ranking, and concept hierarchy. 3) Another attempt is done by (Janet & Reddy, 2011) where the authors aim to improve the analysis and exploration of text-oriented documents; to do so they proposed a new model called Cube Index based on a hierarchical description of each document. This hierarchical description specifies the relationships between words extracted from each document.

4) An AP-structure based on the frequent items named AP-Sets (it comes from Apriori sets, because they are obtained via the Apriori algorithm (Agrawal & Srikant, 1994)) in (Bautista, Molina, Tejeda, & Vila, 2013). This knowledge structure obtained in a constructive way, by initially generating item sets with cardinal equal to 1, next these ones are combined to obtain those of cardinal equal 2, and by continuing until getting item sets of maximal cardinal, with a fixed minimal support. Therefore, the final structure is that of a set of AP-Sets; and 5) Concept hierarchy in CXT-Cube (Oukid, Asfari,

Bentayeb, Benblidia, & Boussaid, 2013) where the semantic dimension extracted from a domain ontology related to the dimension area. The authors applied relevance propagation technique on this hierarchy in order to consider the text semantics during the computation of the concepts weights. Relevance propagation is a technique widely used in Information Retrieval area.

These models allow analyzing the semantics of textual data, while the structural aspect remains untapped. Moreover, conventional multidimensional models are strongly subject to the orthogonality constraint of dimensions that prohibits inter-dimension links, thus leading to a lack of flexibility in design, and then in analysis.

The second category includes new concepts within models. The two well-known models are Galaxy model (Pujolle, Ravat, Teste, & Tournier, 2011) and Complex objects model (Boukraa, Boussaid, Bentayeb, & Zegour, 2011).

The Galaxy model defines the warehouse model as a set of entities where each entity presents as a dimension; a node could link several dimensions when they are compatible for analysis. The main drawback of this work is the authors do not define rules to assist the design phase of a Galaxy model.

The second model, based on Complex object, allows analysis at different granularity levels of each complex object. The complex object model is a three-level model: a class diagram of candidate facts and candidate dimensions represents the first level. In the second level, classes describing the same complex object are grouped into a single package, to provide at the end a diagram of packages describing complex objects. The third level is denoted by a package diagram that results from projecting a complex object package from the second level as an object and associating it with a set of dimension objects described by complex objects related to the object made by complex relationships.

Most of these works use different techniques to manage documents and to incorporate them in a multidimensional model; in addition, the source texts are usually XML documents or texts having an internal structure. To recapitulate, no multidimensional modeling proposal based on structures of XML documents and on semantic contents simultaneously have been made so far. In addition, no work proposes rules or algorithms to assists the Document Warehouse designer elaborating the model: identification of dimensions, hierarchies, etc.

To overcome the drawbacks of the literature works, we have suggested a new multidimensional model called Diamond model that we consider as an extension with semantic aspects to the Galaxy model (Pujolle et al., 2011). In the Diamond model, the semantic is materialized with an appropriate dimension called the Semantic dimension, which is extracted from a semantic resource of the domain of the document set. Our proposed model provides for analysis at different semantic and structural levels. In addition, it allows flexibility to decision-makers since the fact is not predefined; instead, it is selected among dimensions when querying the Diamond model.

3. PROPOSED APPROACH FOR GENERATING XML DOCUMENT WAREHOUSE

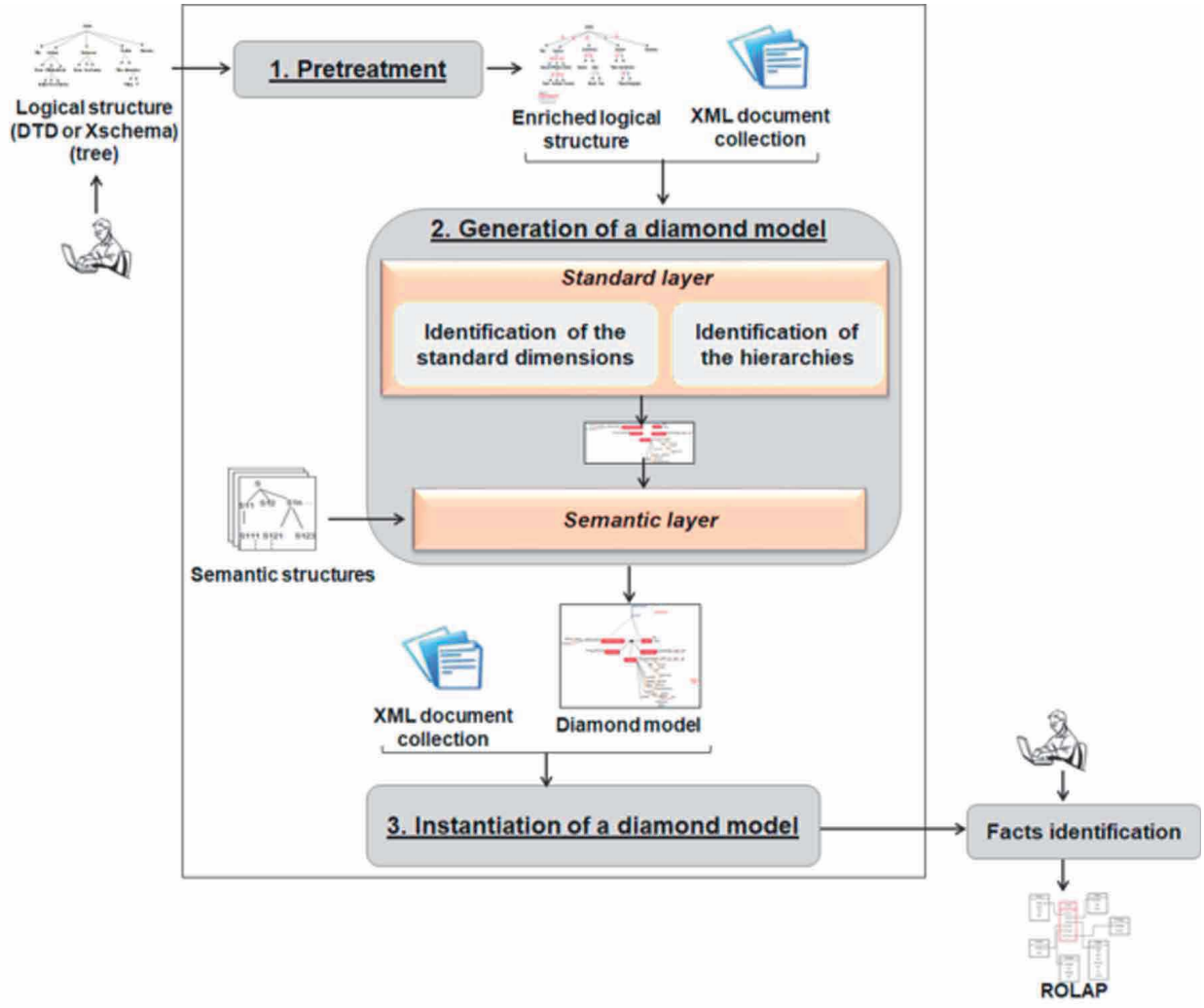
Let us remember that the multidimensional modeling aims at designing multidimensional models that support OLAP analysis. Standard Data Warehouse models are not appropriate for warehousing unstructured data. To alleviate this difficulty, we propose a new multidimensional model for XML document warehousing called Diamond model. The Diamond model is inspired from the Galaxy model in the sense it is built around the single concept of dimension that represents not only an analysis axis but also a plausible subject of analysis. It is characterized by a main extension: the semantic dimension that connects conventional dimensions. Note that linking dimensions of the multidimensional schema is a novel idea we propose.

For the Diamond model, we develop a design approach that produces a Diamond DocW schema from a collection of text-oriented XML documents conform to a logical structure[1] (i.e., DTD or XSchema). We have set in (Azabou, Khrouf, Feki, Soulé-Dupuy, & Vallès, 2015) a three-phase quasi-automatic approach: 10 Pretreatment, 2) Generation of a Diamond multidimensional model, and 3) Instantiation of the multidimensional model (cf. Figure 1). Hereafter we describe these phases shortly.

- **Pretreatment Phase:** The logical structure (i.e., DTD or Xschema) of the input set of documents to warehouse is poor semantically. This phase aims to remedy this issue by injecting a particular semantics. In order to make easier to the user the understanding of the semantics contained in the logical structure, we annotate the links between the DTD elements with their typology. In the logical structure, we distinguish three types of links: 1) Descriptive link: A descendant element represents descriptive information for its parent, 2) Structural link that is Parent-child relationship, and 3) Temporal link: element describing a date as the month, year, etc. The DocW designer should carefully define these elements by examining the content of a representative sample of XML document set because the quality of the annotation influences the quality of the generated Diamond model.
- **Generation of a Diamond Multidimensional Model:** For this phase, we define a set of eleven heuristic rules to generate a Diamond model quasi-automatically, which should include all elements of the enriched logical structure of the input collection of XML documents.
- **Instantiation of the Diamond Model:** This phase consists in automatically detecting the various components of the Diamond multidimensional model (dimensions, and parameters) from the input XML documents.

The next section details the Generation phase of a Diamond model.

Figure 1. Quasi-automatic approach for the generation of a Diamond multidimensional DOCW model



4. MULTIDIMENSIONAL MODEL GENERATION

For readability reasons of the paper, we first introduce the Galaxy model and then our Diamond model, which is an extension of the Galaxy.

The Galaxy model is a network of entities (i.e., dimensions) connected by nodes. Each node denotes compatible entities. Compatible entities can be used together within the same OLAP context (i.e., query). The Galaxy allows flexibility to decision-makers during their specification of multidimensional queries, i.e., without constraining them to select a fact among predefined analysis subjects.

The Galaxy model has a main lack; it considers the structural aspects of the documents, which allow only analysis at different structural levels, while the semantic aspect is completely absent.

To overcome this limitation, our Diamond model focuses on three aspects: 1) textual data structure, 2) textual data semantics, and 3) flexibility of analyses.

The Diamond model consists of two layers:

- **Standard Layer:** It is composed of Standard dimensions. The first level elements of the documents' structure constitute the analysis axes. For each of these elements, its descendants constitute the parameters (organized into hierarchies) or the weak attributes (associated with parameters) (cf. (Azabou et al., 2015) for further details). In the standard layer, a central node associates Compatible Standard dimensions, that is, dimensions that can be used together in the same analytical query.
- **Semantic Layer:** It is materialized with a central dimension called Semantic dimension. It is composed of the hierarchy: Id_Descriptor < Semantic resource. The Semantic dimension is extracted from an external semantic resource related to the domain of the document set. In this layer, we associate semantic elements from the semantic dimension with the textual attributes and descriptor of the standard dimension. The "Described by" relationship in Figure 2 represents this link.

Figure 2 shows the meta-model of the Diamond model. This meta-model describes a set of elements and their characteristics. These elements are either attributes, dimensions or hierarchies. We define two types of dimensions (Standard, and Semantic) as a specialization of the Dimension class. Each dimension has one or many hierarchies. The aggregation called Node_Link links dimensions through a central node. Each hierarchy has a set of parameters. A parameter may have weak attributes. We distinguish two subclasses of parameters: Textual and Numeric parameters. We associate an external semantic resource with the semantic dimension. One semantic resource for a collection of XML documents. The semantic resource contains a set of descriptors. In general, a semantic resource may be an ontology, a taxonomy, a thesaurus or any other kind of resource. In our model, we have elected a thesaurus for two reasons: 1) the absence of domain ontologies appropriate to the decisional process, 2) the availability of well-constructed and standard thesaurus in some domains, as in the medical case.

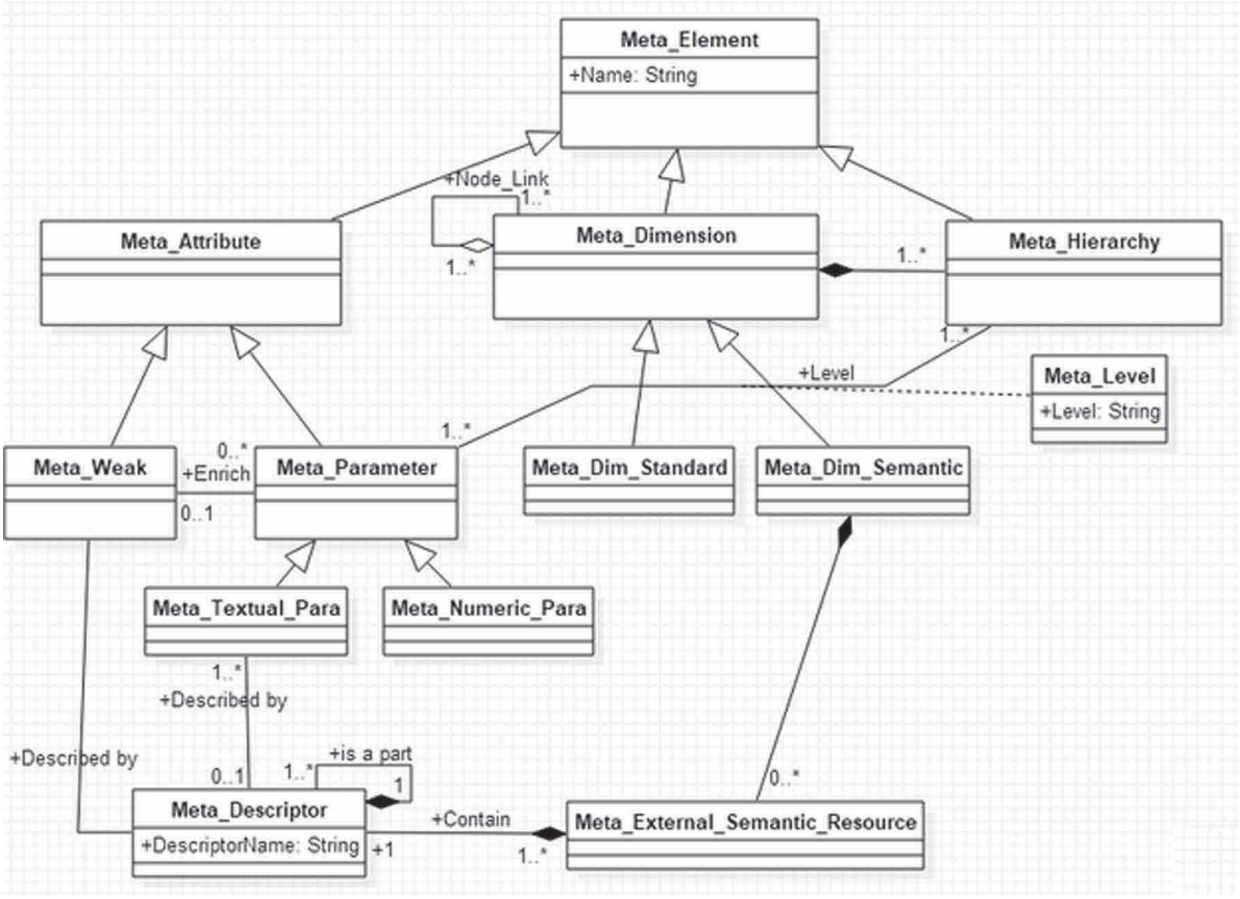
4.1. Standard Layer

In a Diamond model, the standard layer is composed of standard dimensions that may stand for analysis axes or even analysis subjects.

In our approach, we have defined a set of eleven rules in order to help the design of the Diamond model and automate this phase (see Boxes 1-4 and Boxes 6-12). These rules enable to find out the model elements. For clarity reasons, we give these rules (cf. (Azabou et al., 2015) for further details).

Before introducing the different rules, we describe the logical structure of any article description called PubMed and an XML document that conforms to the DTD PubMed (see Figures 9 and 10 in the Appendix).

Figure 2. Meta-model of the Diamond model



4.1.1. Identification of Standard Dimensions

We define the four following rules to determine dimensions from a pretreated tree (cf. Figure 7). In a pretreated tree, we annotate links with letters indicating whether the link is Descriptive, Structural or Temporal. We decide of the letter by exploring XML documents compliant to the logical structure (DTD or XSchema). In Figure 7, we find the two types of links: a Descriptive link: the descendant elements ArticleTitle represent descriptive information for their parent element Article, and a Temporal link the element PubDate gives all the component elements describing the publication date).

Rule 1 identifies the central node that will connect the future dimensions.

4.1.2. Identification of Hierarchies

Once we have extracted the dimensions, the next step identifies the hierarchies. For this, we need to find out the Functional Dependencies (FDs) between the elements of the document structure; we do this by examining a collection of XML documents. Let us remember that in database design, a FD from attribute A to attribute B (A → B), expresses that each value a of A is associated with one, and only one, value b of B (b is not necessarily the same through time; the reverse is not necessarily a FD in general).

Box 1.

Rule 1: The root r of a logical structure S (DTD or XSchema) containing at least one leaf transforms into node for linking future dimensions (to identify using rules later).

Box 2.

Rule 2: Each element d, which is an immediate descendant from the root r and satisfying rule 1 transforms into a dimension named D-d.

Box 3.

Rule 3: Assign an artificial identifier (surrogate key) named Id-D-d to each dimension D-d identified by rule 2. The identifier has rank 1 in the dimension D-d.

Box 4.

Rule 4: For each element d, transformed into a dimension D-d using rule 2, that has no descendant we shall have an attribute (parameter or weak attribute) having the same name D-d as the dimension:
- If the element d contains distinct values then this element becomes a weak attribute, directly connected to the identifier Id-D-d (of dimension D-d).
- Otherwise, it becomes a parameter of rank 2 of dimension D-d (Id-D-d<d).

In multidimensional modeling, a hierarchy organizes parameters from the finest towards the highest granularity. Then there is a FD between two consecutive parameters such as A (e.g. City) is a parameter at a lower level than B (e.g. Country), an immediately higher level to A (City < Country).

Our objective now is to continue extracting the parameters for each dimension already extracted via rule 2. We need heuristic rules that are grouped into three categories depending on the type of link added in the pretreatment phase.

The first category of rules is to process descriptive links and encompasses rules 5 to 9. The second category of rules is to process structural links and has rule 10. Finally, the third category has rule 11 to handle temporal link.

4.1.2.1. Rules for Descriptive Links

A parameter may have associated weak-attributes. A weak attribute is a descriptive information that labels a parameter in OLAP results, it is recommended when the parameter values are meaningless artificial data (as identifiers).

We define five rules (5 to 9) to identify the hierarchies for each element transformed into a dimension (through rule 2) and that has descendants. For these rules, we use the following notation in Box 5:

In some cases, it is necessary to examine a significant collection of XML documents to decide whether an element should transform into parameter or weak attribute. Rule 6 deals with this situation.

Box 5.

- *Ep*: an element of level *p* in a logical structure (the root element has level 1).
- The *k* immediate descendants (i.e. at level *p+1*) for *Ep* are noted a_1, a_2,…a_k.

Box 6.

Rule 5: Each element in a logical structure *S* that has descendants transforms into a multidimensional element of the type *parameter* immediately connected to its father-parameter issued from *S*.

As an illustration, the geographical grouping of cities may differ from country to country. In USA, cities are grouped into countries and there is no departments; however, in France cities are grouped into departments and departments into countries. This semantic issue is left to the warehouse designer who should define two hierarchies:

H-Geog_FR: Id_Client < City < Department < Country
H-Geog_USA: Id_Client < City < Country

4.1.2.2. Rules for Structural Links

Structurally, the elements of a logical structure are organized from the generic toward the specific. We use the Contains symbol "⊃" for describing the organization of the elements; as an example, Section ⊃ Subsection ⊃ Paragraph.

Box 7.

Rule 6: Symmetrical FD
If there are symmetrical FDs between two immediate descendants, a_i and a_j ($i \neq j$) from *Ep* identified as a parameter then:
a. If a_i and a_j may be null-valued then we consider a_i and a_j as two weak attributes associated with *Ep*.
b. If ai may have null values and aj is never null then we consider aj as a parameter immediately connected to parameter *Ep* whereas a_i becomes a weak attribute for a_j.
c. Otherwise (a_i and a_j are never null), a_i and a_j are two consecutive parameters linked to the parameter *Ep.* Arbitrarily, the order of these parameters is $Ep < a_i < a_j$.

Box 8.

Rule 7: No Symmetrical FD
If there is only one FD $a_i \rightarrow a_j$ (without $a_j \rightarrow a_i$) with $i \neq j$ then:
a. If a_i is already a parameter, then a_j transforms into a parameter immediately connected to a_i.
b. If a_i is not yet identified as a parameter then:
- Transform a_i into a parameter.
- Apply case a).

Box 9.

Rule 8: Transitivity within hierarchies
Let a_i, a_j, a_k three attributes, and $a_i \rightarrow a_j$; $a_j \rightarrow a_k$ and $a_i \rightarrow a_k$ three FDs. In the relational database context, the FD $a_i \rightarrow a_k$ is considered as transitive and then should be removed in respect to the non-redundancy database objective. This could not be applied systematically in the data warehouse design. Indeed, $a_i \rightarrow a_k$ may be useful according to the semantic of the application domain. When such a case happens, the designer should decide wether to remove or keep $a_i \rightarrow a_k$. When not removed we obtain two hierarchies defined on the same dimension:
• H1: …$a_i < a_j < a_k$ … and
• H2: …$a_i < a_k$ … (where $a_i < a_k$ denotes parameter a_i is finer than a_k)

Box 10.

Rule 9: No FD between elements
Given *Ep* identified as a parameter with *k* immediate descendants $a_1, a_2, \ldots, a_k$. If there is no FD starting from a_i ($i \in [1..k]$) towards each of its brothers, then a_i becomes a parameter linked to the parameter *Ep*.

Note that the structural organization order is the inverse to the organization of dimensional hierarchies. To alleviate this problem, we reverse the order of the structural hierarchy elements through using rule 10.

For the elements linked by a structural link (parent-child relationship) in the logical structure, we first apply six rules of descriptive links (rules 5-9) to determine parameters and then we reorganize them by applying rule 10 in order to rectify the structural hierarchy.

4.1.2.3. Rules for Temporal Links

A DW considers data as chronological series of values; therefore, the temporal dimension is systematically present (Kimball & Ross, 2003). Thus, in order to define the temporal dimension, we detect first Date elements in the logical structure and then organize them; this organization relies on a standard Date dimension.

The following rule deals with elements that describe temporal components.

So far, rules 5- 11 deal only with the automatic generation of the standard dimensions from the logical structure. Furthermore, we need to add the Semantic dimension, which reflects the semantics of the unstructured textual elements. We introduce the Semantic dimension in the next section.

4.2. Semantic Layer

A central dimension called semantic dimension composes the semantic layer. This semantic dimension has two parameters (Semantic resource and Id_Descriptor) organized in a hierarchical manner. Id_Descriptor parameter is associated to a weak attribute Descriptor that represents complementary information. In our Diamond model, we have elected a thesaurus MeSH as Semantic resource. MeSH[2] consists of sets of terms naming descriptors in hierarchical structure that permits searching at various levels of specificity[3]. The Id_Descriptor parameter represents the different levels of a Semantic resource the number of Descriptors and levels varies from a Semantic resource to another. For the representation of this dimension, we will use a new type of parameters, named recursive parameter because the semantic resource is represented in a hierarchical manner.

Each level $l_i = < d_1, d_2, ...d_n >$ includes a set of descriptors d_j ($j \in [1,n]$) extracted from a Semantic resource. Figure 3 depicts an example of a Semantic resource MeSH.

The Semantic dimension allows the user to enrich the analysis focusing on the meaning behind the textual data in addition to the lexical information (set of terms).

The main idea of a Diamond model is to explore the content of the text documents in the database. In order to achieve this, we associate semantics to the textual content of documents through semantic links between each textual attribute belonging to a standard dimension and its appropriate descriptor in the semantic dimension. Thus, we can perform analysis by the abstract of the article or by descriptors related to the abstract (for example, 'Neoplasms', 'Cysts', 'Carcinogens' descriptors).The use of the

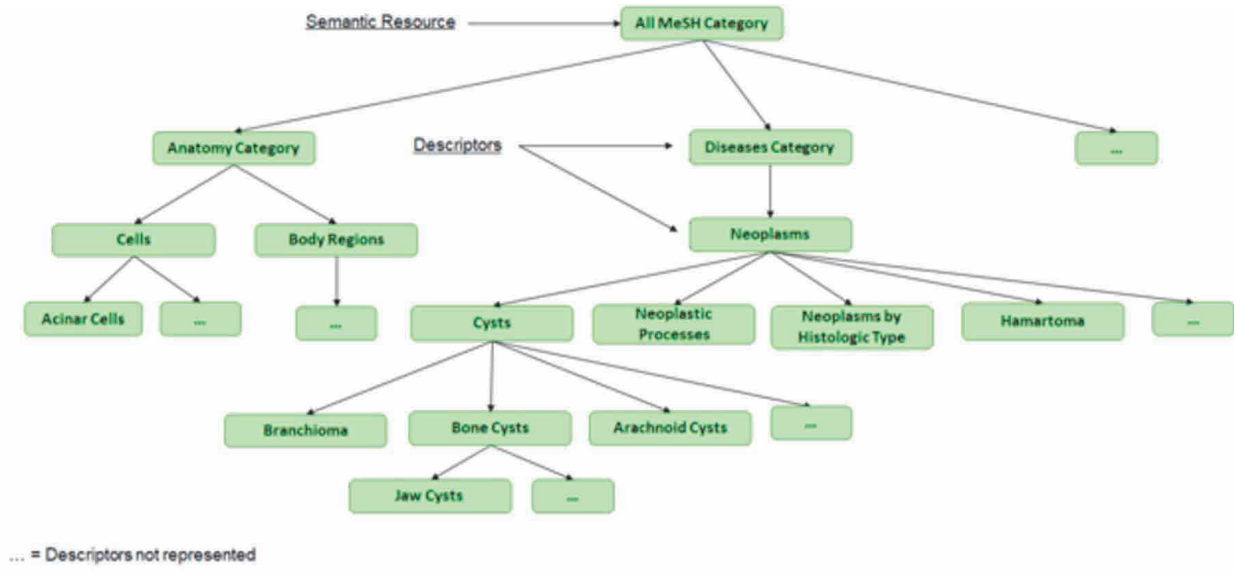
Box 11.

Rule 10: Structural order rule
For the elements linked by a structural link and transformed into parameters identified by the descriptive rules 5 to 9, we inverse the order of these parameters such as the last one becomes the first parameter.

Box 12.

Rule 11: Temporal rule
The set of descendant elements at a same level that describes *temporal* components (e.g., month and day that compose a date) builds a temporal dimension (where each component element becomes a parameter in the hierarchy).

Figure 3. Example of a sample Semantic Resource with descriptors



multidimensional model descriptors enrich the OLAP analysis so the user can introduce the semantic of textual attribute in the queries over the data cube.

The determination of the descriptor for each textual attribute is evaluated according to the approach described in (Ben Mefteh, Khrouf, Feki, & Soulé-Dupuy, 2013) (Ben Mefteh, Khrouf, Feki, & Soulé-Dupuy, 2014). This approach determine a semantic structure for an XML document relying on the logical structure and the content of the document. It includes the four following steps: 1) Extraction of significant terms from leaf elements of document (leaves of the tree structure of the DTD or XSchema). 2) Choice of a thesaurus describing the semantics of a document, 3) Associate descriptors of the selected thesaurus with the leaf elements of the document (descriptors that best reflect the semantics of the terms describing leaf elements); and 4) Inference of descriptors for non-leaf elements. Figure 4 depicts an example of a logical structure and a semantic structure for an XML document.

5. EXPERIMENTS

In order to substantiate our approach, we have implemented a software tool named Diamond-Gen for the generation of a Diamond multidimensional model. It receives as input a collection of XML documents and their logical structure (cf. Appendix) and then produces a Diamond model.

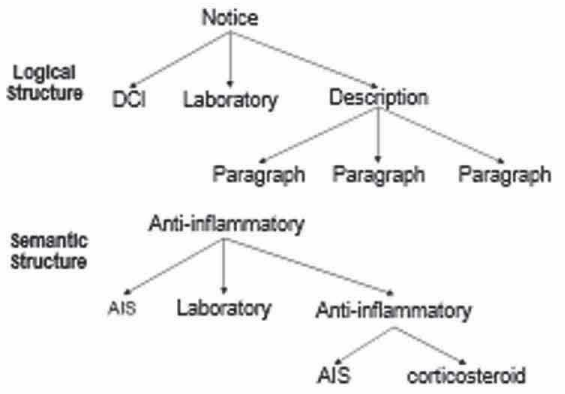
More precisely, we have carried out an experiment using a set of 1500 XML documents taken from the medical collection PubMed[4]. PubMed comprises more than 26 million citations for biomedical literature from MEDLINE[5], life science journals, and online books.

In the remaining of this section, we present the features of our software tool Diamond-Gen while exemplifying them through the Diamond model generated for this experiment.

First, the pretreatment step is launched to produce the enriched DTD for the PubMed DTD. Figure 7 depicts the input DTD enriched with the typology of links between the DTD elements annotated automatically. We find the two types of links: Descriptive (e.g., the Article element has elements ArticleTitle and Language), and Temporal (e.g., the element PubDate gives all the component elements describing the publication date).

Secondly, dimensions and their hierarchies of the Diamond model are extracted by applying the eleven rules defined in Section 4. Thus, a Semantic dimension extracted from a Semantic Resource.

Figure 4. Example of a logical structure and its associated semantic structure ((Ben Mefteh et al., 2014))



We associate the descriptors from the semantic dimension with the textual attributes of the standard dimension

Figure 8 illustrates the obtained Diamond model for this example. It is composed of the following five dimensions: D-PMID, D-DateCreated, D-Article, D-MedlineJournalInfos and D-Keywordlist.

In order to assess the result of the prototype, we have compared the Diamond generated Galaxy model with the model built manually by an expert; the result of this comparison is encouraging because the two models are very close.

For example, an analyst wants to analyze the number of publications by Title of journal and Year of creation (cf. Figure 5). With a simple multidimensional table, he can obtain the number of publication according to a set of terms representing Title of journal and Year of creation. In order to present, the semantic of the Title of journal it is necessary to enrich the analysis focusing on the meaning behind the Title of journal (cf. Figure 6).

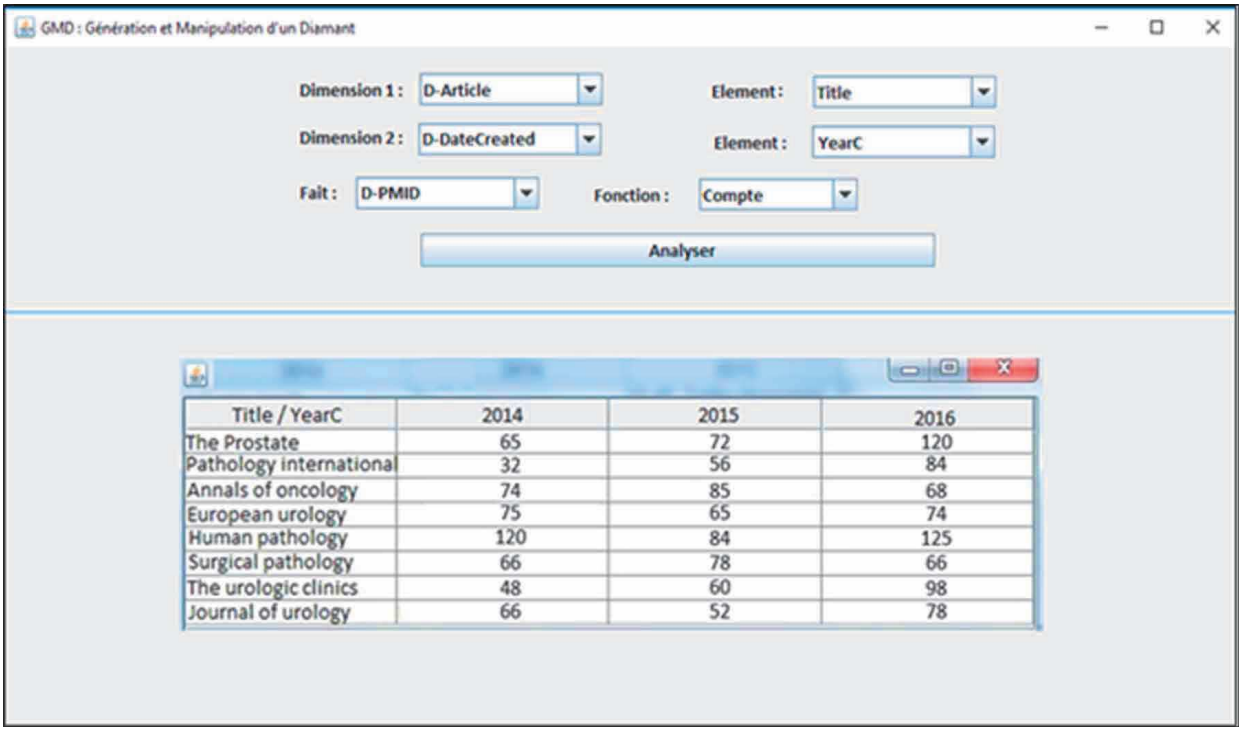
Besides the factual data, the decision-maker wants to analyze the number of documents by the topics approached for each title of journal and year of creation. The classic OLAP tools do not allow assisting the decision-maker in this query because of the absence of operators and methods adapted to the management of textual data. To overcome this limit, we propose a new operator called To_Semantic. To_Semantic allows the passage from the textual attribute (i.e. title of journal) of a standard dimension to the descriptor of a semantic dimension that describes it. This operator takes as input a multidimensional table having at least one textual attribute in one of its dimensions and generates a new multidimensional table based on the descriptors of the MeSH thesaurus (see Box 13).

As an example, we use the To_Semantic operator to analyze the number of publications by the topics for each title of journal and year (cf. Figure 6).

6. CONCLUSION

The document warehouse technology did not pay enough attention to exploring documents in the decisional process despite these documents hold valuable information as important as numeric data extracted from the operational systems. Surely, XML documents constitute an important source of information for decision-makers. Therefore, they largely merit to be warehoused according to specific multidimensional models and then analyzed using appropriate OLAP operators that take into consideration structural and semantic aspects of documents. In this paper, we have proposed a dedicated approach to build a multidimensional model for documents: it is the Diamond model. This model mainly consists of two layers: standard layer (a set of standard dimensions) constructed from the structure of a set of XML documents, and a Semantic layer (a semantic dimension). The main objective of the Semantic dimension is to switch from the simple text to a semantic level.

Figure 5. A sample multidimensional table (the number of publication according to a Title of journal and a Year of creation)

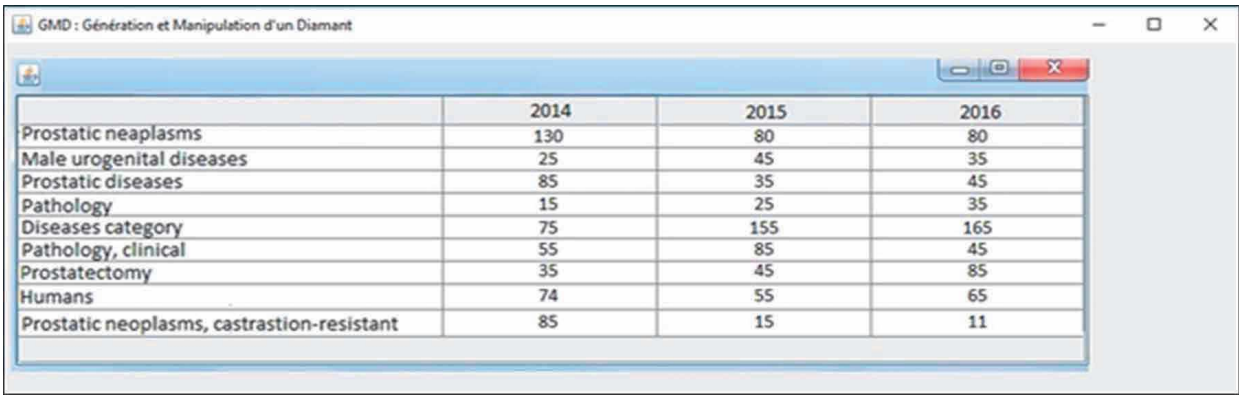


Title / YearC	2014	2015	2016
The Prostate	65	72	120
Pathology international	32	56	84
Annals of oncology	74	85	68
European urology	75	65	74
Human pathology	120	84	125
Surgical pathology	66	78	66
The urologic clinics	48	60	98
Journal of urology	66	52	78

Box 13.

TM_{RES}= To_Semantic (TM_{IN}, D_i, Att)
- **TM_{IN}**: Multidimensionnel table
- **D ∈ {D1, D2}**: One of the two dimensions displayed in MT_{IN} have a textual attribute.
- **Att**: A textual attribute of D (parameter or weak attribute) displayed in TM_{IN}

Figure 6. Multidimensional analysis of publications counts displayed by the semantic of the Title of journal and by years



	2014	2015	2016
Prostatic neaplasms	130	80	80
Male urogenital diseases	25	45	35
Prostatic diseases	85	35	45
Pathology	15	25	35
Diseases category	75	155	165
Pathology, clinical	55	85	45
Prostatectomy	35	45	85
Humans	74	55	65
Prostatic neoplasms, castrastion-resistant	85	15	11

Likewise, we have presented a software tool, called Diamond-Gen that implements the method for the generation of Diamond model. We performed an experimental study on a PubMed collection. For this experiment, we obtained a Diamond with five dimensions.

As an immediate extension for this work, we plan to evaluate Diamond-Gen software tool on more XML structures. In addition, we are in the step of finishing the definition of a set of analytical operators based on semantic dimension that take into consideration the specificities of the Diamond

Figure 7. PubMed DTD enriched

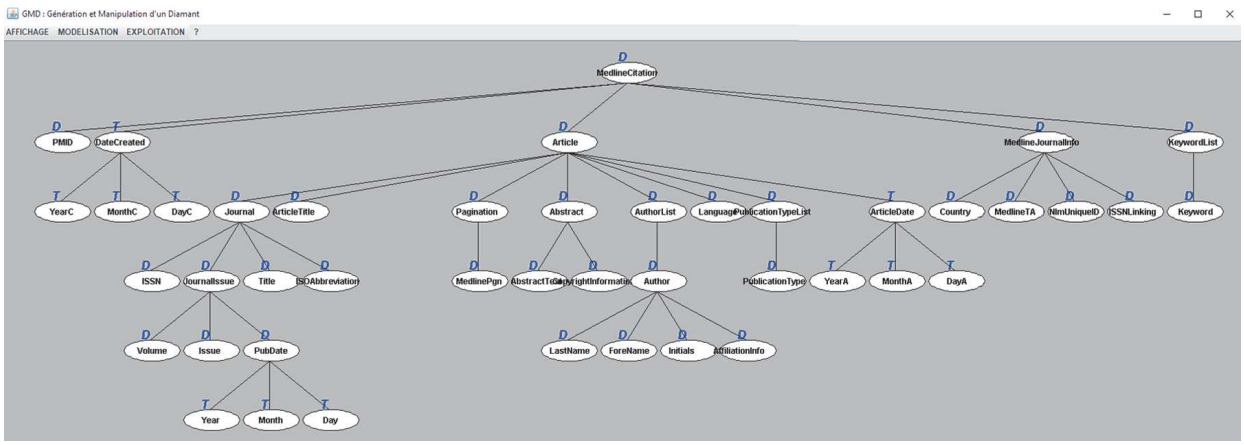
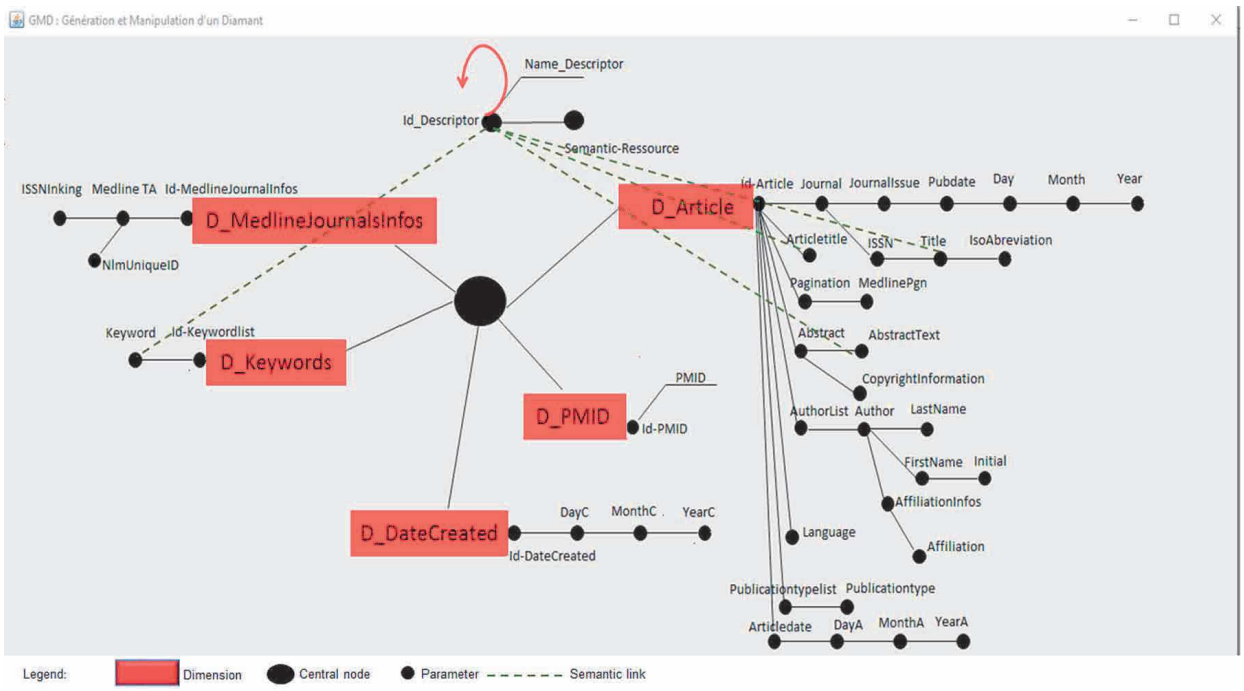


Figure 8. Result Diamond model for for the PubMed collection



model. They will facilitate the interpretation of the results of the multidimensional analyses on the textual data. Finally, we plan to evaluate the scalability of our approach.

Frequently organizations handle large volumes of unstructured data. Relevant examples are emails, forum discussions and documentation. Faced with this increasing production of data, it becomes difficult to store, interrogate, model and analyze this very Big volumes of data. Therefore, it would be interesting to evaluate the architecture of a document warehouse in a Big Data context.

REFERENCES

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for mining Association rules. In *Proceedings of VLDB*, Santiago, Chile (pp. 487-499).

Azabou, M., Khrouf, K., Feki, J., Soulé-Dupuy, C., & Vallès, N. (2014). A Novel Multidimensional Model for the OLAP on Documents: Modeling. In *Generation and Implementation. International Conference on Model & Data Engineering MEDI'2014*, Larnaca, Cyprus, September 24-26 (pp. 258-272).

Azabou, M., Khrouf, K., Feki, J., Soulé-Dupuy, C., & Vallès, N. (2015). Diamond multidimensional model and aggregation operators for document OLAP. In *IEEE Ninth International Conference on Research Challenges in Information Science (RCIS)* (pp. 363-373). doi:10.1109/RCIS.2015.7128897

Azabou, M., Khrouf, K., Feki, J., Soulé-Dupuy, C., & Vallès, N. (2016). Analyzing textual documents with new OLAP operators. In *13th IEEE/ACS International Conference of Computer Systems and Applications (AICCSA)* (pp. 1-8).

Bautista, M., Molina, C., Tejeda, E., & Vila, A. (2013). A new multidimensional model with text dimensions: definition and implementation. In *International Conference IPMU*, Dortmund, Germany (pp. 158-167).

Ben Mefteh, S., Khrouf, K., Feki, J., & Soulé-Dupuy, C. (2013). Semantic Structure for XML Documents: Structuring and pruning. *Journal of Information Organization*, *3*(1), (pp. 36–46).

Ben Mefteh, S., Khrouf, K. Feki, J., & Soulé-Dupuy, C. (2014) Structuration sémantique des documents XML: Expérimentations et évaluation. In *CORIA-CIFED* (pp. 61-70).

Boukraa, D., Boussaid, O., Bentayeb, F., & Zegour, D. (2011). Modèle multidimensionnel d'objets complexes du modèle d'objets aux cubes d'objets complexes. *Ingénierie des Systèmes d'Information*, *16*(6), (pp. 41–65). doi:10.3166/isi.16.6.41-65

Golfarelli, M., Maio, D., & Rizzi, S. (1998). The Dimensional Fact Model: A Conceptual Model for Data Warehouses. *International Journal of Cooperative Information Systems*, *7*(2-3), (pp. 215–247).

Immon, W. H. (2005). *Building the Data Warehouse*. John Wiley and sons.

Janet, B., & Reddy, A. V. (2011) Cube Index for Unstructured Text Analysis and Mining. In *ICCC'11 Proceedings of the 2011 International Conference on Communication, Computing & Security* (pp. 397-402). doi:10.1145/1947940.1948023

Kimball, R., & Ross, M. (2003). *The Data Warehouse Toolkit*. New York: Wiley.

Lin, C. X., Ding, B., Han, J., Zhu, F., & Zhao, B. (2008). Text cube: Computing in measures for multidimensional text database analysis. In *Eighth IEEE International Conference on Data Mining* (pp. 905-910).

Oukid, L., Asfari, O., Bentayeb, F., Benblidia, N., & Boussaid, O. (2013). CXT-cube: contextual text cube model and aggregation operator for text OLAP. In *International Workshop On Data Warehousing and OLAP (DOLAP)*, San Francisco, CA (pp. 27-32). doi:10.1145/2513190.2513201

Pujolle, G., Ravat, F., Teste, O., & Tournier, R. (2011). Multidimensional Database Design from Document Centric XML Documents. *In DAWAK'11: International Conference on Data Warehousing and Knowledge Discovery*, France (pp. 51-65). doi:10.1007/978-3-642-23544-3_5

Yu, Y., Lin, C., Sun, Y., Chen, C., Han, J., Liao, B., & Zhao, B. et al. (2009). iNextCube: Information network-enhanced text cube. In *VLDB'09: Proceedings of the 35th International Conference on very Large Data Bases*, Lyon, France (Vol. 2, pp. 1622-1625).

Zhang, D., Zhai, C., & Han, J. (2009). Topic cube: Topic modeling for olap on multidimensional text databases. In *SDM '09: Proceedings of the 2009 SIAM International Conference on Data Mining*, Sparks, NV (pp. 1124–1135). doi:10.1137/1.9781611972795.96

ENDNOTES

1 A logical structure is a tree of tags representing an XML document.
2 MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity
3 http://slideplayer.com/slide/4237273
4 Http://www.ncbi.nlm.nih.gov/PubMed
5 Https://www.nlm.nih.gov/bsd/pmresources.html

APPENDIX

Figure 9. PubMed DTD

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT MedlineCitation (PMID, DateCreated, Article, MedlineJournalInfo, KeywordList*)>
<!ELEMENT Abstract (AbstractText,CopyrightInformation)>
<!ELEMENT AbstractText (#PCDATA)>
<!ELEMENT Affiliation (#PCDATA)>
<!ELEMENT Article (Journal,ArticleTitle,Pagination,Abstract, AuthorList, Language,PublicationTypeList, ArticleDate*)>
<!ELEMENT ArticleDate (YearA,MonthA,DayA)>
<!ELEMENT ArticleTitle (#PCDATA)>
<!ELEMENT Author (LastName, ForeName, Initials, (AffiliationInfo))>
<!ELEMENT AuthorList (Author)>
<!ELEMENT AffiliationInfo (Affiliation)>
<!ELEMENT CopyrightInformation (#PCDATA)>
<!ELEMENT Country (#PCDATA)>
<!ELEMENT DateCreated (YearC,MonthC,DayC)>
<!ELEMENT DayC (#PCDATA)>
<!ELEMENT DayA (#PCDATA)>
<!ELEMENT Day (#PCDATA)>
<!ELEMENT ForeName (#PCDATA)>
<!ELEMENT ISOAbbreviation (#PCDATA)>
<!ELEMENT ISSN (#PCDATA)>
<!ELEMENT  ISSNLinking (#PCDATA)>
<!ELEMENT Initials (#PCDATA)>
<!ELEMENT Issue (#PCDATA)>
<!ELEMENT Journal (ISSN, JournalIssue, Title, ISOAbbreviation)>
<!ELEMENT JournalIssue (Volume, Issue, PubDate)>
<!ELEMENT Keyword (#PCDATA)>
<!ELEMENT KeywordList (Keyword)>
<!ELEMENT Language (#PCDATA)>
<!ELEMENT LastName (#PCDATA)>
<!ELEMENT MedlineJournalInfo (Country, MedlineTA, NlmUniqueID,ISSNLinking)>
<!ELEMENT  NlmUniqueID (#PCDATA)>
<!ELEMENT  MedlinePgn (#PCDATA)>
<!ELEMENT Month (#PCDATA)>
<!ELEMENT MonthC (#PCDATA)>
<!ELEMENT MonthA (#PCDATA)>
<!ELEMENT PMID (#PCDATA)>
<!ELEMENT Pagination (MedlinePgn)>
<!ELEMENT PubDate (Year,Month, Day)>
<!ELEMENT PublicationType (#PCDATA)>
<!ELEMENT PublicationTypeList (PublicationType)>
<!ELEMENT Title (#PCDATA)>
<!ELEMENT Volume (#PCDATA)>
<!ELEMENT Year (#PCDATA)>
<!ELEMENT YearA (#PCDATA)>
<!ELEMENT YearC (#PCDATA)>
```

Figure 10. An XML document that conforms to the PubMed DTD

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE MedlineCitation SYSTEM "D:\programme maha\DTD\DTDpubmed1.dtd">

<MedlineCitation Owner="NLM" Status="In-Data-Review">
  <PMID Version="1">27037211</PMID>
  <DateCreated>
    <YearC>2016</YearC>
    <MonthC>04</MonthC>
    <DayC>23</DayC>
  </DateCreated>
  <Article PubModel="Print-Electronic">
    <Journal>
      <ISSN IssnType="Electronic">1879-3061</ISSN>
      <JournalIssue CitedMedium="Internet">
        <Volume>27</Volume>
        <Issue>5</Issue>
        <PubDate>
          <Year>2016</Year>
          <Month>May</Month>
        </PubDate>
      </JournalIssue>
      <Title>Trends in endocrinology and metabolism: TEM</Title>
      <ISOAbbreviation>Trends Endocrinol. Metab.</ISOAbbreviation>
    </Journal>
    <ArticleTitle>Genomic Aberrations Drive Clonal Evolution of Neuroendocrine Tumors.</ArticleTitle>
    <Pagination>
      <MedlinePgn>242-4</MedlinePgn>
    </Pagination>
    <Abstract>
      <AbstractText Label="METHODS">Molecular features of castration-resistant neuroendocrine prostate
      <CopyrightInformation>Copyright © 2016 Elsevier Ltd. All rights reserved.</CopyrightInformation>
    </Abstract>
    <AuthorList CompleteYN="Y">
      <Author ValidYN="Y">
        <LastName>Kaushik</LastName>
        <ForeName>Akash Kumar</ForeName>
        <Initials>AK</Initials>
        <AffiliationInfo>
          <Affiliation>Verna and Marrs McLean Department of Biochemistry and Molecular Biology and Alk
        </AffiliationInfo>
      </Author>
      <Author ValidYN="Y">
        <LastName>Sreekumar</LastName>
        <ForeName>Arun</ForeName>
        <Initials>A</Initials>
        <AffiliationInfo>
          <Affiliation>Verna and Marrs McLean Department of Biochemistry and Molecular Biology and Alk
        </AffiliationInfo>
      </Author>
    </AuthorList>
    <Language>eng</Language>
    <PublicationTypeList>
      <PublicationType UI="D016428">Journal Article</PublicationType>
    </PublicationTypeList>
    <ArticleDate DateType="Electronic">
      <YearA>2016</YearA>
      <MonthA>03</MonthA>
      <DayA>29</DayA>
    </ArticleDate>
  </Article>
  <MedlineJournalInfo>
    <Country>United States</Country>
    <MedlineTA>Trends Endocrinol Metab</MedlineTA>
    <NlmUniqueID>9001516</NlmUniqueID>
    <ISSNLinking>1043-2760</ISSNLinking>
  </MedlineJournalInfo>
  <KeywordList Owner="NOTNLM">
    <Keyword MajorTopicYN="N">DNA methylation</Keyword>
    <Keyword MajorTopicYN="N">adenocarcinoma</Keyword>
    <Keyword MajorTopicYN="N">genomic aberrations</Keyword>
    <Keyword MajorTopicYN="N">molecular biomarker</Keyword>
    <Keyword MajorTopicYN="N">neuroendocrine</Keyword>
    <Keyword MajorTopicYN="N">prostate cancer</Keyword>
  </KeywordList>
</MedlineCitation>
```

Kais Khrouf obtained his doctorate in Computer Science from the University of Toulouse III, France. He is currently an Assistant Professor at the Jouf University. His research focuses on information Systems, data warehousing, OLAP and Conceptual modeling. He is also a member of the MIRACL research laboratory, Sfax, Tunisia.

Jamel Feki is Professor in Computer Science at the Faculty of Economics and Management of the University of Sfax in Tunisia. He has been working at the University of Jeddah, in Saudi Arabia, since January 2015. His main research area is Data Warehousing and Mining; he has supervised several PhD theses and published more than 100 papers in specialized journals and conferences. He is a permanent member of the Miracl research Laboratory and is a member of several international conferences.

Chantal Soulé-Dupuy is Full Professor in Computer Science (PhD in 1990) at the Faculty of Information technology of the University of Toulouse Capitole. She was the Dean of the Faculty of Information Technology for 10 years and is currently the co-director of the Doctoral School on Mathematics, Computer Science and Telecommunications of Toulouse. Her current research interests include data modelling and analysis, information retrieval, structured and unstructured information indexing and warehousing, information integration and annotation. She also has a strong experience in managing large-scale repositories of heterogeneous information in a variety of application areas (including biology, health, aviation and space). She has published more than 100 papers in specialized journals and conferences and is a member of several international conference committees.

Nathalie Vallès is currently Assistant Professor in the University of Toulouse France. She has published more than 50 papers in specialized journals and conferences and is a member of several international conferences.